Towards Understanding Educational Technology Interventions with a Pareto Efficiency Perspective

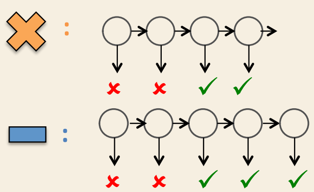
José González-Brenes, Pearson

PEARSON Research & Innovation NETWORK

Yun Huang, University of Pittsburgh

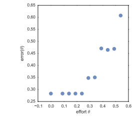
EDUCATIONAL INTERVENTIONS

- Educational interventions have a cost (effort) to the learner, and a pay-off (outcome)
- Human-propelled machine learning interventions are evaluated with Randomized control trials ($$$) or with classification evaluation metrics
- For example: Adaptive tutoring systems minimize student practice, and maximize their outcomes. Optimizing them independently is trivial (E.g, don't teach at all, or teach for 100 years each concept).
- Adaptive tutoring systems are evaluated on how predictive they are on future student performance



LEARNER EFFORT-OUTCOME PARADIGM (LEOPARD)

- Effort: how much practice the tutor gives to the student
- Outcome: how well does the student does after tutoring / Error: 1- Outcome
- White (Whole Intelligent Tutoring System Evaluation) metric that operationalizes Leopard. Drop-in replacement for Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. Extends work from Lee & Brunskill (2012)
- Problem? ill-specified models are not concave

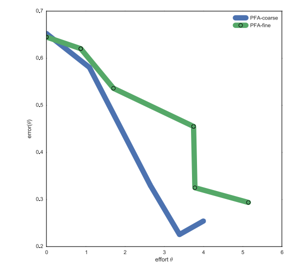


- Counterfactual simulation of what the tutor would have done
- Varying thresholds gives a Pareto frontier

predicted performance

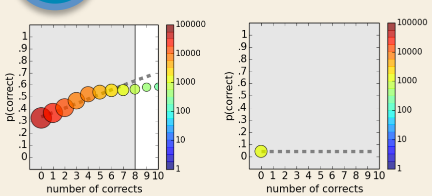
actual performance

	t	student u	skill q	$\hat{c}_{u,q,t+1}$	$y_{u,q,t}$	
effort=	0	Alice	s1	**.6**		
	1	Alice	s1	.5	0	
	2	Alice	s1	.5	1	2/3 score
	3	Alice	s1	.6	1	
	0	Bob	s1	.4		
effort=	1	Bob	s1	**.7**	1	
	2	Bob	s1	.7	1	
	3	Bob	s1	.7	1	
	4	Bob	s1	.8	0	4/5 score
	4	Bob	s1	.9	1	
	6	Bob	s1	.9	1	



FOUR QUESTIONS YOU SHOULD ASK YOURSELF ABOUT THE VALIDITY OF YOUR EVALUATION

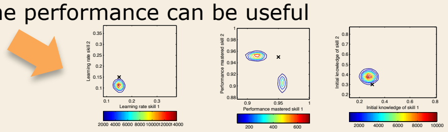
1 Your model is accurate - but is it useful?



	AUC	F	score	effort
Bad student model	**.85**	**.79**	.18	10.1
Baseline	.50	0	.18	11.2

We trained a "bad student model" with real student data with flat learning curves. The model is very accurate, yet is not useful for adaptivity. Solutions:

- Report classification accuracy averaged over skills (for models with 1 skill per item)
  ✖ Not useful for comparing or discovering different skill models
- Report as "difficulty" baseline
  ✖ Experiments suggest that models with baseline performance can be useful
- Use Leopard



2 Suboptimal decisions?

Cognitive model	AUC	score	effort
Coarse (27 skills)	.69	.41	**55.73**
Fine (90 skills)	.74	.36	**88.16**

The fine model gives 50% more of practice to students – yet it has better AUC.

3 Unstable results?

Yudelson and Ritter '2015 demonstrated that a change of 0.01 RMSE can have a a HUGE change in tutoring policies

4 What are you measuring?

Simulations using synthetic data suggest that classification evaluation metrics have low correlation to what we typically would measure with a RCT